Proceedings of Coling: The 20th International Conference on Computational Linguistics, Geneva 23-27 August 2004, 31-38

Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health

Barry SMITH
Department of Philosophy
University at Buffalo
Buffalo, NY 14260, USA
and
Institute for Formal Ontology
and Medical Information Science
Saarland University
D-66041 Saarbrücken, Germany
phismith@buffalo.edu

Christiane FELLBAUM
Department of Psychology
Princeton University
Princeton, NJ 08544, USA
and
Berlin-Brandenburg Academy of Sciences
Berlin, Germany
fellbaum@princeton.edu

Abstract

A consumer health information system must be able to comprehend both expert and non-expert medical vocabulary and to map between the two. We describe an ongoing project to create a new lexical database called Medical WordNet (MWN), consisting of medically relevant terms used by and intelligible to non-expert subjects and supplemented by a corpus of natural-language sentences that is designed to provide medically validated contexts for MWN terms. The corpus derives primarily from online health information sources targeted to consumers, and involves two sub-corpora, called Medical FactNet (MFN) and Medical BeliefNet (MBN), respectively. The former consists of statements accredited as *true* on the basis of a rigorous process of validation, the latter of statements which non-experts *believe to be true*. We summarize the MWN / MFN / MBN project, and describe some of its applications.

1 From WordNet to Medical WordNet

WordNet is the principal lexical database used in natural language processing (NLP) research and applications. (Miller, 1995), (Fellbaum, ed., 1998) While WordNet's current version (2.0) has broad medical coverage, it manifests a number of defects, which reflect both the lack of domain expertise on the part of the responsible lexicographers, and also the fact that WordNet was not built for domain-specific applications. The research community has long been aware of these defects (Magnini and Strapparava, 2001), (Bodenreider and Burgun, 2002), (Burgun and Bodenreider, 2001), (Bodenreider, *et al*., 2003). Our response is to create Medical WordNet (MWN), a free-standing lexical database designed specifically for the needs of natural-language processing in the medical domain, with the goal of removing the 'noise' which is associated with the application of WordNet and similar resources to this specialized domain.

MWN's initial focus is on English single-word expressions as used and understood by non-experts. We systematically review WordNet's existing medical coverage by assembling a validated corpus of sentences involving specific medically relevant vocabulary. Input to our validation process includes the definitions of medical terms already existing in WordNet, and also sentences generated via the semantic relations linking such terms in WordNet. In addition, input includes sentences derived from online medical information services targeted to consumers.

Our methodology is designed (1) to document natural language sentential contexts for each relevant word sense in such a way that the expressed information can be (2) validated by medical experts and (3) accessed automatically by NLP applications such as information retrieval, machine translation, question-answer systems, and text summarization.

A major stumbling block for existing NLP applications is automatic sense disambiguation. An automatic system can detect with high

reliability that a given occurrence of a word like *feel* or *dead* is a verb or adjective. But it cannot easily determine which of a variety of alternative meanings such polysemous words have in any given context.

WordNet's architecture, designed for representing and distinguishing word senses, has made an important contribution towards a solution of the automatic word sense disambiguation problem. Our corpus of English language sentences relating to medical phenomena is designed to build upon this contribution. The corpus is restricted to grammatically complete, syntactically simple sentences in natural language which have been rated as *understandable* by non-expert human subjects in controlled questionnaire-based experiments. It is restricted in addition to sentences which are self-contained in the sense that they make no reference to any prior context and do not contain any proper names, or anaphoric elements (like *it* or *he* or *then*) that need to be interpreted with respect to other sentences or some surrounding discourse or context. This corpus is designed to be used initially for purposes of quality assurance of MWN and also to support the population of MWN by yielding new families of words and word senses for inclusion. As will become clear, however, our use of human validators will allow us to extend the usefulness of the corpus in a variety of ways. Thus we can use it to build new sorts of applications for information retrieval in the domain of consumer health. But it also allows new avenues of research in linguistics and psychology, for example in allowing us to explore individual and group differences in medical knowledge and vocabulary, and in understanding non-expert medical reasoning and decision-making.

2 Medical FactNet and Medical BeliefNet

To this end, however, we need to exploit our validation data to create two sentential subcorpora, called Medical FactNet (MFN) and Medical BeliefNet (MBN), respectively.

MFN consists of those sentences in the corpus which receive high marks for *correctness* on being assessed by medical experts. MFN is thus designed to constitute a representative fraction of the *true beliefs* about medical phenomena which are intelligible to non-expert English-speakers. MBN consists of those sentences in the corpus which receive high marks for *assent.* MBN is thus designed to constitute a representative fraction of the *beliefs* about medical phenomena (both true and false) distributed through the population of English speakers.

The validation process that is involved in the construction of MFN is used to detect errors in the existing WordNet, and also to ensure that the coverage of the natural language medical lexicon in MWN is of a scientific level sufficient to allow MWN technology to work in tandem with terminology and ontology systems designed for use by experts.

Both MFN and MBN inherit from MWN the formal architecture of the Princeton WordNet. (Fellbaum, ed., 1998) However, we enhance this architecture to maximize its usefulness in medical information retrieval.

Compiling MFN and MBN in tandem allows systematic assessment of the disparity between lay beliefs and vocabulary as concerns medical phenomena and the corresponding expert medical knowledge. The ultimate goal of our work on MFN is to document the entirety of the medical knowledge that can be understood by average adult consumers of healthcare services in the United States today. If the methodology for the creation and validation of the corpus here described proves successful, then we believe that the preconditions for the realization of this much larger goal will have been established. Responses from NLP researchers and from online information providers to our initial work on MFN/MBN convinces us that this realization would have considerable significance for the management and retrieval of consumer health information in the future.

3 Background and Motivation

Recent studies of the use of computer-based tools for consumer health information retrieval point to a mismatch between existing tools and the non-expert language used by most consumers – the language used not only by patients but also by family members, advisors, administrators, lawyers, and so forth, and to some degree also by nurses and physicians. (Slaughter, 2002), (C. A. Smith, *et al*., 2002), (Tse, 2003), (Tse and Soergel, 2003), (McCray and Tse, 2003), (Zeng, *et al*., in press)

Where the usage of medical terms by professionals is at least in principle subject to control by standardization efforts, the highly contextually dependent usage of medical terms on the part of lay persons is much more difficult to capture in applications – and this in spite of the fact that it is in many ways simpler than expert usage. The taxonomies reflecting popular lexicalizations in all domains are indeed much less elaborate at both the upper and lower levels than in the corresponding technical lexica. (Medin and Atran, eds., 1999) Thus there are no popular

terms linking *infectious disease* and *mumps*, so that in the popular medical taxonomy of diseases the former immediately subsumes the latter. The popular medical vocabulary naturally covers only a small segment of the encyclopedic vocabulary of medical professionalsm, and it lexicalizes mainly at the level of taxonomic orders. Popular medical terms (*flu*) are often fuzzier than technical medical terms. Many popular terms also cover a larger range of referent types than do technical terms; others may cover only part of the extension of their technical counterparts. We hypothesize, however, that with few exceptions the *focal* meanings (Berlin and Kay, 1969) of expert and non-expert terms will be identical. Constructing MFN and MBN allows us to test this and related hypotheses in a systematic way.

4 Mismatches in Doctor-Patient Communication

The skills of a physician in general practice comprise the ability to acquire relevant and reliable information through communication with patients through the use of non-expert language and to convey diagnostic and therapeutic information in ways tailored to the individual patient.

Since the physician, too, is a member of the wider community of non-experts, and continues to use the non-expert language for everyday purposes, one might assume that there are no difficulties in principle keeping him from being able to formulate medical knowledge in a vocabulary that the patient can understand. As (Slaughter, 2002) and (C. A. Smith, *et al.*, 2002) have shown, however, there are limits to this competence. The former examines dialogue between physicians and patients in the form of question-answer pairs, focusing especially on the relations documented in the UMLS Semantic Network. Only some 30% of the relations used by professionals in their answers directly match the relations used by consumers in formulating their questions. An example of one such question-answer pair is taken from (Slaughter, p. 224):

Question Text: My seven-year-old son developed a rash today that I believe to be chickenpox. My concern is that a friend of mine had her 10-day-old baby at my home last evening before we were aware of the illness. My son had no contact with the infant, as he was in bed during the visit, but I have read that chickenpox is contagious up to two days prior to the actual rash. *Is there cause for concern at this point?*

Answer Text: (a) Chickenpox is the common name for varicella infection. [...] (b) You are correct in that a person with chickenpox can be contagious for 48 hours before the first vesicle is seen. [...] (c) The fact that your son did not come in close contact with the infant means he most likely did not transmit the virus. (d) Of concern, though, is the fact that newborns are at higher risk of complications of varicella, including pneumonia. [...] (e) There is a very effective means to prevent infection after exposure. A form of antibody to varicella called varicella-zoster immune globulin (VZIG) can be given up to 48 hours after exposure and still prevent disease.

Such examples illustrate also that there are lexically rooted mismatches in communication (which may in part reflect legal and ethical considerations) between experts and non-experts. Professionals often do not re-use the concepts and relations made explicit in the questions put to them by consumers. In our example, the questioner requests a yes/no-judgment on the possibility of contagion in a 10-day-old baby. In fact, however, only section (c) of the answer responds to this question, and this in a way which involves multiple departures from the type of non-expert language which the questioner can be presumed to understand. Rather, physicians expand the range of concepts and relations addressed (for example through discussion of issues of prevention, etc.).

In all cases, the information source, whether it be a primary care physician or an online information system, must respond primarily with *generic information* (i.e. with information that is independent of case or context), and this is so even where requests relate to specific and episodic phenomena (occurrences of pain, fever, reactions to drugs, etc.). (Patel, *et al.*, 2002) In our example, all sections except for (c) are of this generic kind. They contain answers in the form of context-independent statements about causality, about types of persons or diseases, about typical or possible courses of a disease. MFN is accordingly designed to map the generic medical information which non-experts are able to understand.

5 Non-Expert Language in Online Communication

Understanding patients requires both explicit medical knowledge and tacit linguistic competence dispersed across large numbers of more or less isolated practitioners. This is not a problem so long as this knowledge is to be applied locally, in face-to-face communication with patients. However, as a result of recent developments in technology, including telemedicine and internet-based medical query systems, we now face a situation where such dispersed, practical (human) knowledge does not suffice.

(Ely, *et al.*, 2000) and (Jacquemart and Zweigenbaum, 2003) have shown that clinical

questions are expressed in a small number of different syntactic-semantic patterns (about 60 patterns account for 90% of the questions). Such yes/no questions are of the forms: *Do hair dyes cause cancer?*, *Can I use aspirin to treat a hangover?* With the right sort of information resource, questions such as these can easily be transformed automatically into statements providing correct answers: *Hair dyes can cause bladder cancer*, *Aspirin doesn't help in case of a hangover* , and these answers can be linked further to relevant and authoritative sources.

Query text	**MEDLINEplus® response** (with links to documents sorted by the following keywords)
tremor	**Tremor**, Multiple Sclerosis, Parkinson's Disease, Degenerative Nerve Diseases, Movement Disorders
intentional tremor	**Tremor**, Multiple Sclerosis, Parkinson's Disease, Spinal Muscular Atrophy, Degenerative Nerve Diseases
tremble	Anxiety, Parkinson's Disease, Panic Disorder, Caffeine, **Tremor**
trembling	Anxiety, Parkinson's Disease, Panic Disorder, Phobias, **Tremor**

Table 1: Online-Inquiry to MEDLINEplus®
(http://www.nlm.nih.gov/medlineplus)

MEDLINEplus is described in its online documentation as a source of medical information for both experts and non-experts 'that is authoritative and up to date.' Enquirers can use MEDLINEplus like a dictionary, choosing health topics by keywords. Alternatively, they can use the system's search feature to gain access to a database of relevant online documents selected for reliability and accessibility on the basis of pre-established criteria.

Table 1 shows the problems that can arise when a system fails to take account of the special features of the knowledge and vocabulary of typical non-expert users. Here success in finding the needed information depends too narrowly on the precise formulation of the query text. Thus *tremble* and *trembling* call forth different responses (one lists caffeine, the other phobias), even though the terms in question differ only in a morphological affix that does not involve a meaning distinction. Such problems are characteristic of information services of this kind. Experienced internet users are of course familiar with the limitations of search engines, and so they are able to manipulate their query texts in order to get more and better results. Even experienced users, however, will not be able to overcome the arbitrary sensitivities of an information system, and the latter cannot have the goal of bringing non-experts' ways of using language into line with that of the system.

6 Corpus- and Fact-Based Approaches to Information Retrieval

(Patel, *et al*., 2002) make clear that if a medical information system is to mediate between experts and non-experts, then it must rest on an understanding of both expert and non-expert medical vocabulary. But terms, or word forms, are not always associated with word meanings in a clear-cut and unambiguous fashion; and the problem of polysemy is compounded when different speaker populations are involved. A lexical database must represent all and only the meanings of each given term in such a way that these meanings can be clearly discriminated and mapped onto word occurrences in natural text and speech. Achieving these ends is one of the hardest challenges facing both theoretical and applied linguistic science today. It is generally agreed that the meanings of highly polysemous terms cannot be discriminated without consideration of their contexts (e.g., Pustejovsky, 1995). People manage polysemy without apparent difficulties; but modeling human speakers' capacity for lexical disambiguation in automatic language processing tasks is hard. The idea underlying the present proposal draws on currently emerging NLP methodologies that harness the ability of powerful and fast computers to store and manipulate both lexical databases and large collections of text collections or corpora. The strategy is to train automatic systems on large numbers of semantically annotated *sentences* that are naturally used and understood by human beings, and to exploit standard pattern-recognition and statistical techniques for purposes of disambiguation. Words and the representation of their senses, stored in lexical databases, can be linked for this purpose to specific occurrences in corpora.

7 Related Work

Currently, several resources are being built in the spirit of this methodology. Examples are FrameNet (Baker, *et al.*, 1998), (Baker, *et al*., 2003) and Penn Proposition Bank (Kingsbury and Palmer, 2002), both of which focus on word usage in general, rather than on domain-specific contexts. In contrast to our own project, neither of the mentioned resources attempts to build a corpus in a systematic way that is designed to ensure adequate coverage of some given domain. Furthermore, neither project is concerned with the questions of factuality or validation of statements.

Another project with goals similar to those of

MFN is the CYC (short for enCYClopedia) knowledge base, a collection of hundreds of thousands of statements, mostly about the external world, such as: *The earth is round, Albany is the capital of New York.* (Lenat, 1995), (Guha, *et al*., 1990) These statements, which were entered over many years by CYC employees, are parcelled out into separate micro-theories devoted to different domains. (On CYCs medical coverage see (Bodenreider and Burgun, in press).)

Our work differs in a number of ways from CYC: (i) we focus on one single (albeit very large) domain; (ii) CYC does not store English sentences but rather – in keeping with its goal of being language-unspecific – statements couched in the symbolism of a modified first-order logic; (iii) CYC incorporates folk beliefs and expert knowledge indiscriminately, and its separate micro-theories are not designed to be consistent either with each other or with the body of established science; (iv) only a reduced part of CYC is publicly available.

8 WordNet

WordNet 2.0 is a large electronic lexical database of English that has found wide acceptance in areas as diverse as artificial intelligence, natural language processing, and psychology. (Agirre and Martinez, 2000), (Al-Halimi and Kazman, 1998), (Artale, *et al*., 1997), (Basili *et al*., 1997), (Berwick, *et al*., 1990), (Burg and van de Riet 1998), (Cucchiarelli and Velardi 1997), (Fellbaum 1990), (Gonzalo, *et al*., 1998), (Harabagiu and Moldovan, 1996) Its coverage, which is comparable to that of a collegiate dictionary, extends over some 130,000 word forms. The most common application is in information technology, where it is used for information retrieval, document classification, question-answer systems, language generation, and machine translation. WordNet was originally conceived as a full-scale model of human semantic organization, and its design was guided by early experiments in artificial intelligence. (Collins and Quillian, 1969)

WordNet was quickly embraced by the NLP community, a development that has guided its subsequent growth and design, and WordNet is now widely recognized as the lexical database of choice for NLP. The appeal of WordNet's design is reflected in the fact that wordnets have been, and continue to be, built in dozens of languages. Wordnets supporting many European and non-European languages are already available. All are linked to the original English WordNet, which thereby functions as an interlingual index. In consequence, all wordnets can be mapped to one another. This means that our work on Medical WordNet will ultimately be translatable into dozens of languages with very little additional effort.

8.1 Architecture of WordNet

The building blocks of WordNet are synonym sets ('synsets'), which are unordered sets of distinct word forms and which correspond closely to what, in medical terminology research, are called 'concepts.' Membership in a synset requires that the word forms express the same concept and are in this sense 'cognitively synonymous' (Cruse, 1986). More formally, synset members must be interchangeable in some sentential contexts without altering the truth-value of the sentences involved. WordNet's architecture is thus grounded in the notion of *truth-preserving interchangeability of word forms in sentential contexts*, although research has not thus far focused on this feature. Constructing Medical FactNet allows us to rectify this gap by making explicit the contexts in which word forms are used in an environment that allows the systematic testing of the effects of word form substitution.

Examples of synsets are {*car, automobile*} or {*shut, close*}. WordNet 2.0 contains some 115,000 synsets, with many word forms belonging to a plurality of synsets.

WordNet is a *net* in virtue of the fact that the synsets are linked to one another via a small number of binary relations that differ for each of the four syntactic categories covered by WordNet: nouns, adjectives, verbs, adverbs. Noun synsets are interlinked by means of the subtype (or *is-a*) relation, as exemplified by the pair *poodle-dog*, and by means of the *part-of* relation, as exemplified by the pair *tire-car*. Verb synsets are connected by a variety of lexical entailment relations that express manner elaborations, temporal relations, and causation (*walk-limp*, *forget-know*, *show-see*). (Fellbaum, 2002), (Fellbaum, 2003) Thus if X limps, then X also walks, but not *vice versa.* The links among the synsets structure the noun and verb lexica into hierarchies, with noun hierarchies being considerably deeper than those for verbs.

WordNet's appeal for NLP applications stems from the fact that its synset architecture can be exploited in building NLP applications that target the problem of automatic word sense disambiguation. Although most word forms in English are monosemous (*clinician*, *epidemic*), the most frequently occurring words are highly polysemous (*host*, *dress*, *arm*). The ambiguity of a polysemous word in a context can be resolved by distinguishing the multiple senses in terms of their links to other words within the WordNet *net*. For

example, the noun *club* can be disambiguated by an automatic system that considers the superordinates of the different synsets in which this word form occurs: *association*, *playing card*, and *stick*.

The information contained in WordNet is of two sorts: lexical (i.e. verbal) knowledge, stored in WordNet's synset architecture, and encyclopedic (i.e. factual) knowledge, found in the definitions (or 'glosses') associated with each concept. These definitions can be problematic, however, as they were generated by lexicographers who were not specialists in the domains to which the words in the synsets belong. Often, the definitions were modelled on those found in existing dictionaries, but in these cases, too, problems have arisen above all in the form of a mismatch between definitions representing technical (specialist) knowledge and definitions reflecting non-expert usage. To resolve this problem each synset in MWN is augmented with two glosses. One is formulated for the layman, the other in expert language.

A further problem turns on the fact that the sentences included in WordNet 2.0 as illustrations of the use of synonyms in sentential contexts do not always reflect correct or characteristic usages of the words in the synset. Constructing MFN addresses this problem in a systematic way.

9 Uses of WordNet in Medical Informatics

(Xiao and Rösner, 2003) shows how WordNet can be used as a tool for simplifying information extraction from MEDLINE. Parsing tools are used to extract verbs from the corpus of MEDLINE abstracts, and it is then shown that very many (both low- and high-frequency) verbs are grouped together into WordNet synsets in such that, within this specific discourse domain, there is only one conceptual relation linking all the verbs in each of the relevant synsets. In this way it is possible to simplify the process of information abstraction by reducing the number of relations that need to be taken into account in the analysis of texts.

(Buitelaar and Sacaleanu, 2001) describe work showing how, using the German version of WordNet, one can use statistical analysis to support automatic selection of the most likely synset associated with given nouns appearing in medical corpora.

WordNet's design allows users with specific technical applications to augment the database, primarily by adding new terms as leaves to the existing branches of its subsumption and part-whole hierarchies. Such enriched wordnets retain all of the original information, and the added words are semantically specified in terms of WordNet's relations. (Turcato, *et al*., 2001) and (Buitelaar and Sacaleanu, 2002) describe an attempt to extend the German wordnet with synsets pertaining to the medical domain using automatic methods, in particular the detection of semantic similarity from co-occurrence patterns in a domain-specific corpus. The results, while good, are hampered by problems of lexical polysemy and by the characteristically German tendency for compound formation, which leads to potentially open-ended lexicon growth, and thus poses posing great problems for automatic word sense recognition and discrimination. One clear conclusion from this study is that fully automated lexical acquisition provides inadequate results, and that much of the work must be performed manually. Our proposal reflects this conclusion.

(Bodenreider and Burgun, 2002) and (Burgun and Bodenreider, 2001) characterize the definitions of anatomical concepts in WordNet and in various portions of the UMLS Metathesaurus. They found that anatomical definitions are characteristically of the form: superordinate + distinguishing feature (the latter expressed through some form of adjectival modification or relative clause, etc.). This way of defining words is in fact the canonical one (for nouns, and, to some degree, for verbs as well) and lexicographers follow it as much as possible when writing definitions. MWN will observe this standard consistently in its augmentation and standardization of WordNet's definitions, drawing on the results of the studies of best practice in the formulation of definitions in biomedical terminologies and ontologies in (Smith and Rosse, 2004), (Bodenreider, *et al*., 2004) and (Smith, *et al*., 2004).

10 The Medical Coverage of WordNet 2.0

For the verb feel, WordNet 2.0 distinguishes in all 13 separate meanings, of which at least the following have an obvious medical significance, and are handled by WordNet in rough accordance with their usage in medical contexts:

3. **sense** – (perceive by a physical sensation, e.g., coming from the skin or muscles: *He felt the wind; She felt an object brushing her arm; He felt his flesh crawl; She felt the heat when she got out of the car; He feels pain when he puts pressure on his knee.*)

4. **feel** – (seem with respect to a given sensation given: *My cold is gone – I feel fine today; She felt tired after the long hike*)

10. **palpate, feel** – (examine (a body part) by palpation: *The nurse palpated the patient's stomach; The runner felt her pulse*)

For the adjective *dead*, WordNet 2.0 distinguishes 21 meanings, with only two approximating to meanings of this term as used in medical contexts:

1. **dead** (vs. alive) – (no longer having or seeming to have or expecting to have life: *The nerve is dead*; *A dead pallor*)
9. **dead**, deadened – (devoid of physical sensation; numb: *his gums were dead from the Novocain*)

Not only does WordNet fail to distinguish those medically relevant meaning distinctions illustrated by phrases such as *dead tissue*, *dead organ*, *dead matter*, *dead cell*, *dead body*, etc., but its definition of the primary medically relevant sense of *dead* (as: 'no longer having or seeming to have or expecting to have life') runs together three separate notions which it is medically important to keep separate.

WordNet recently added domain labels to many synsets. One such label is **medicine**; others are **surgery** and **drug**. However, it was left undecided on what criteria terms should be selected as domain labels and what the relations among the relevant domains should be (arguably, **surgery** and **drug** should be included in the wider domain of **medicine**). In addition, labels were not systematically assigned to WordNet terms. Currently, when asked to output terms associated with *medicine*, the browser returns some 504 nouns, verbs, and adjectives (both single words and phrases), representing some 270 different senses. On the other hand, many cognate senses with clear medical uses are currently *not* labeled in this way. Table 2 provides examples, with the *medicine* label picked out in bold:

autopsy #1	{autopsy, necropsy, postmortem, PM, postmortem examination – (an examination and dissection of a dead body to determine cause of death or the changes produced by disease)}
fester #1	{fester, maturate, suppurate – (ripen and generate pus; *her wounds are festering*)}
festering #1	{festering, suppuration, maturation – (**(medicine)** the formation of morbific matter in an abscess or a vesicle and the discharge of pus)}
festering #2	{pus, purulence, suppuration, ichor, sanies, festering – (a fluid product of inflammation)}
infection #1	{(the pathological state resulting from the invasion of the body by pathogenic microorganisms)}
infection #3	{(**(medicine)** the invasion of the body by pathogenic microorganisms and their multiplication which can lead to tissue damage and disease)}
infection #4	{infection, contagion, transmission – (an incident in which an infectious disease is transmitted)}
maturation #2	{growth, growing, maturation, development, ontogeny, ontogenesis – ((biology) the process of an individual organism growing organically; a purely biological unfolding of events involved in an organism changing gradually from a simple to a more complex level; *he proposed an indicator of osseous development in children*)}
maturation #3	{festering, suppuration, maturation – (**(medicine)** the formation of morbific matter in an abscess or a vesicle and the discharge of pus)}
zymosis #2	{(**(medicine)** the development and spread of an infectious disease (especially one caused by a fungus))}

Table 2. Examples of Medically Relevant Entries in WordNet 2.0

Table 2 also illustrates the degree to which WordNet currently includes obsolete medical terms (*ichor*, *morbific*, *unction*) and also terms drawn seemingly indiscriminately from both technical medical vocabularies and from natural language. Some synsets contain only folk or only technical terms, some contain a mixture of both. Definitions are largely taken over from medical dictionaries prepared for experts.

To provide a preliminary estimate of the extent of WordNet's somewhat arbitrary medical coverage we derived a test lexicon of 2838 single-word medical terms from an existing digitalized lexical resource for medical language processing (LinKBase of the Belgian NLP company L&C), which was constructed independently of WordNet by medical professionals. The method used was to transform LinKBase into an alphabetically ordered term list and to eliminate automatically all acronyms, all multi-word terms, all proprietary terms, all terms containing numbers, and all terms longer than 10 letters. Remaining technical terms were then removed manually. Of the residual 2838 terms, only 11 were not present in any form

in WordNet 2.0, though considerably more were not treated adequately in regard to their specifically medical usages. Almost all missing terms were compounds such as *bedwetting, breastfed, coldsore*.

WordNet 2.0 has inadequate treatment of the systematic polysemy of nouns like *dizziness* and *itching*. These, like many other nouns, are both sensations and symptoms. The symptom role is also not encoded for many other nouns, including *redness*, *retching*, *swelling*, and so forth. WordNet states: *a tumor is a mass of tissue* and *a tumor is abnormal*, but not: *some tumors are malignant*.

WordNet's treatment of *is-a*, *part-of* and other relations, too, is marked by inadequacies in the medical domain. Thus WordNet currently contains a verb entailment relation exemplified by the pair *snore-sleep* defined as: 'if someone snores, then he necessarily also sleeps.' In medicine, however, it is recognized that it is quite possible to snore while awake, since snoring is there defined as the respiratory induced vibration of glottal tissues and this is associated not only (and most usually) with sleep but also with relaxation or obesity.

Our methodology for constructing MFN involves the validation by experts of all relations between WordNet's medically relevant synsets. It provides us with a systematic means to detect such errors. Constructing MBN gives us in addition the resources to do justice to the reason why such cases were included in WordNet in the first place: *People can only snore when they are asleep* and similar sentences belong precisely to the folk beliefs about medicine which MBN documents – not, however, to MFN. More generally, constructing MBN in tandem with MFN allows us to highlight those cases where non-experts and experts use the same term in different ways.

Another family of terms currently poorly treated in WordNet are those manifesting polysemy along the medical/non-medical axis. For example, the medical senses of *recession* or *rigors* are not recorded in WordNet 2.0. A lexical database for purposes of automatic sense disambiguation must clearly differentiate all such senses. (Computerized medical information systems do not offer the possibility of follow-up in the sort of cases of misunderstanding which we have in communication between laypersons and medical practitioners.) Thus while MWN will contain only word *forms* that are used by non-experts (and thus part of natural rather than technical language), it must for practical reasons record word *senses* that are peculiar to the technical vocabulary.

11 Method for Translating Online Content into Basic Sentences

We carried out experiments designed to test a variety of methodologies for deriving terms and sentences for our corpus, including elicitation experiments with expert and non-expert human subjects, and data-mining from online bulletin boards. We established that the most promising sources for both term- and sentence-generation are certain online information sources targeted specifically to non-specialist users.

In one experiment the basic sentences meeting our MFN/MBN criteria were derived by researchers in medical informatics from factsheets on *Airborne allergens* in NIAID's Health Information Publications and on *Hay fever and perennial allergic rhinitis* in the UK NetDoctor's Diseases Encyclopedia.

There is no good way to tell the difference between allergy symptoms of runny nose, coughing, and sneezing and cold symptoms. Allergy symptoms, however, may last longer than cold symptoms. *from NIAID HealthInfo* (*information also included in MEDLINEplus*)	1. Allergies have symptoms. 2. Colds have symptoms. 3. A runny nose is a symptom of an allergy. 4. Coughing is a symptom of an allergy. 5. Sneezing is a symptom of an allergy. 6. Cold symptoms are similar to allergy symptoms. 7. A cold is not an allergy. 8. Allergy symptoms may last longer than cold symptoms.
What is hay fever? Hay fever, otherwise known as seasonal allergic rhinitis, is an allergic reaction to airborne substances such as pollen that get into the upper respiratory passages – the nose, sinus, throat – and also the eyes. *from NetDoctor* (UK)	1. Hay fever is an allergy. 2. Hay fever is an allergic reaction. 3. Hay fever is a type of allergy. 4. Hay fever is a type of allergic reaction. 5. Hay fever is a reaction to pollen. 6. Hay favor is a reaction to airborne substances. 7. In hay fever airborne substances get into the nose. 8. In hay fever airborne substances get into the throat. 9. In hay fever airborne substances get into the eyes.

Table 3: Sample sentences derived from online medical information sources

The initial documents were divided into paragraph-length sections, and raters were instructed to associate with each section complete self-contained sentences expressing the generic medical knowledge it contains. Sentences were to be formed using simple syntax and as far as possible drawing on terms used in the original

sources. Processors were instructed to eliminate sentences containing anaphora, indexical expressions, formulations of instructions, warnings and the like, and to replace them where possible by complete statements constructed via simple syntactic modifications. Subjects were instructed to include only such terms and information which they themselves judged would be intelligible to non-experts.

1644 sentences were produced, representing some 20 person hours of effort; examples are presented in Table 3. 500 of these sentences were subjected to a preliminary evaluation, each sentence being presented to pairs of beginning medical students for independent evaluation. 58% of the sentences were rated by both members of each pair with a score of 5. However, a closer analysis of the results revealed that the weighted kappa measure for inter-rater agreement was too low for these results to be statistically significant. Further testing of this methodology will thus call for larger sample sizes and for the use of raters with specific expertise in relation to the phenomena described.

12 Sources and Selection

The primary sources for terms in MWN and for sentences in our test corpus are the relevant general lexical information contained in WordNet, supplemented by medical dictionaries and large medical terminology and ontology systems such as MeSH and LinKBase, and by internet resources such as MEDLINEPlus and NetDoctor focusing especially on **coverage of common diseases**. We shall maintain an internet portal through which links to sources used and the results of our term- and sentence-extraction will be made available online as raw data for use by other researchers.

In this initial phase of our project we are interested primarily in self-contained generic (case- and context-independent) statements with a relatively simple syntax. To derive such sentences we use two methods:

Method 1 derives sentences from a lexical database such as WordNet. We treat the database as a set of links between terms of the form *t*L*u* (where L ranges over 'is-a', 'part-of', and other relations) and *t*, *u* range over terms which occur in the medical sublexicon. Some members of the resulting class of *t*L*u* formulas can be transformed automatically into English sentences with a minimal amount of post-processing. For example each '*t is-a u*' formula can be transformed into sentences of the forms 'a *t* is a *u*' and 'a *t* is a type of *u*' (with corrections for articles and plurals, as in: *A cut is a type of wound; An abrasion is a wound; Patients are people*). Others must be subject to manual extraction, which can be carried out by native English-speakers (linguists or others trained in manipulation of lexical databases) with no special medical expertise. Each extracted sentence is given a precise identifying number and associated with metadata identifying its source.

Method 2 derives single sentences from on-line consumer health information sources along the lines described in section 11 above. Here each sentence in the source documentation is given a precise identifying number, indicating source document, position in this document, and section from which sentences have been inferred. Extracted sentences, too, are given precise identifying numbers and are associated with metadata documenting section and document of origin, date of processing, and also individual responsible for extraction.

13 Human Subject Validations

The output sentences from the above will serve, together with a random infusion of non-medical, folk-medical-but-false and medical-but-technical sentences, as inputs to validations carried out by human subjects. These will be of three primary types, referred to in what follows as U, B and C, for *understanding, belief,* and *correctness*, respectively. All sentences will pass through the U filter, in which laypersons will be recruited to rate sentences for *understandability*. Those sentences which survive will pass on to B and C. In B laypersons will rate sentences for *degree of belief*, in C medically trained participants ('experts') will rate sentences for correctness. Statements receiving a rating of 4 or higher (out of a range from 1 to 5) from each of two raters in B will be stored as components of Medical BeliefNet; statements receiving a similarly high rating from each of two raters in C will be stored as components of Medical FactNet. The ratings for all sentences, both those which do and those which do not pass through the MBN/MFN filters, will be stored for further analysis.

14 Future Work

We envisage the MBN/MFN methodology being used in the fields of medical education and medical literacy to evaluate the reliability of the medical knowledge of different non-expert communities. On the basis of metadata pertaining to the sources of entries in MBN it will be possible to keep track of specific kinds of false beliefs as originating in specific populations of informants. This may prove a valuable source of information in targeting particular groups for specific types of remedial medical education.

Furthermore, the extended MBN will provide opportunities for a new type of research in the field of consumer health. Specifically, we envisage experiments that investigate how the domain of medical phenomena is conceptualized by non-expert human subjects. Cognitive psychologists and anthropologists such as Rosch and others (Rosch, 1975), (Rosch, 1973), (Rosch, 1978) have postulated a level of lexical specification that they call 'basic level.' Basic level words correspond to *basic kinds* in the ontology of language-using subjects. Such words exist in all semantic domains, but they have been studied predominantly among words denoting natural kinds, such as animals, vegetables, fruit, and colors. (Medin and Atran, 1999), (Berlin and Kay, 1969) For example, *tomato* is often cited as an example of a basic level word, whereas *vegetable* is a superordinate, and *cherry tomato* is a subordinate. Basic level words have many striking properties: they are universally lexicalized, characterized by high frequency of occurrence, and they are learned first by children. The concepts they denote have properties that differ maximally from each other (e.g., a tomato is very different from a cabbage or a bean), but the difference between a basic level word and a subordinate (such as between a tomato and a cherry tomato) is less pronounced. The basic level lexicon in the medical domain has thus far not been explored, but such research promises important theoretical benefits. MBN might be used to determine the basic level in the domain under investigation by examining the difference in the frequency of occurrence of synonyms: highly frequent terms are good candidates for basic level words. We can then use the results of this work to provide a specification of the non-expert ontology of the medical domain and begin to explore differences between it and the ontologies underlying expert medical terminologies.

Note that, in all of the above, MFN and MBN have characteristically played different roles. Thus where MFN has been associated with constructing practical tools designed to assist users in coming to believe what is true, MBN has been associated with research, for example regarding what people believe about medical phenomena.

15 Towards a Comprehensive Documentation of Consumer Health Knowledge

We estimate that the two documents referred to in Table 3 above represent together some 0.5 % of the information available on these two sites that is relevant to the purposes of constructing a comprehensive survey of consumer health knowledge. This suggests that a future comprehensive version of MFN might contain some 320,000 sentences. The prospect of constructing a sentence-based information resource of this size would until very recently have rightly been considered overwhelming. The success of WordNet gives us confidence that this problem, too, can be solved.

16 Acknowledgements

Work on this paper was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation. Smith's work was further supported by the European Union Network of Excellence on Semantic Datamining, and by the project "Forms of Life", sponsored by the Volkswagen Foundation. Our thanks go in addition to Werner Ceusters, Christopher Chute, James Cimino, Jean-Pierre Koenig, David Mark, Daniel Osherson, and Martin Trautwein.

References

Agirre, E., Martinez, D. Exploring automatic word sense disambiguation with decision lists and the Web. *Proceedings of the Semantic Annotation and Intelligent Annotation Workshop*, organized by COLING, Luxembourg, 2000.

Al-Halimi, R., Kazman, R. Temporal indexing through lexical chaining. Fellbaum C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Maryland, May 1998.

Artale, A., Magnini, B., Strapparava C. WordNet for Italian and its use for lexical discrimination. *Proceedings of the 5th Congresso dell' Associazione Italiana per l'Intelligenza Artificiale*, Rome, September 1997; 16-19.

Baker, C. F., Fillmore, C. J., Cronin, B. The structure of the framenet database. *International Journal of Lexicography*, 2003; 16.3: 281-296.

Baker, C. F., Fillmore, C. J., Lowe, J. B. The Berkeley FrameNet project. *Proceedings of the COLING-ACL*, Montreal, Canada, 1998.

Basili, R., DellaRocca, M., Pazienza, M. T. Contextual word sense tuning and disambiguation. *Applied Artificial Intelligence* 1997; 11 (3): 235-262.

Berlin, B., Kay, P. *Basic color terms*. Berkeley/Los Angeles: University of California Press, 1969.

Berwick, R., Fellbaum, C., Gross, D., Miller, G. WordNet: A lexical database organized on psycholinguistic principles. In Zernik U. (ed.), *Using On-line Resources to Build a Lexicon*. Erlbaum, Hillsdale, NJ, Erlbaum, 1990; Chapter 9: 211-231.

Bodenreider, O., Burgun, A., Mitchell, J. A. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. *Studies in Health Technology and Informatics* 2003; 95: 379-384.

Bodenreider, O., Burgun, A. Characterizing the definitions of anatomical concepts in WordNet and specialized sources. *Proceedings of the First Global WordNet Conference*, January 2002; 223-230.

Bodenreider, O., Burgun, A. Ontologies in the biomedical domain. Part II: examples. *Journal of the American Medical Informatics Association* (in press).

Bodenreider, O., Smith B., Kumar A., Burgun A. Investigating Subsumption in DL-based terminologies: A case study in Snomed-CT. In: R. Cornet and S. Schulz (eds.), *Proceedings of KR-MED 2004: Knowledge Representation in Medicine*. (Lecture Notes in Bioinformatics 2994), Springer, Berlin: 2004.

Buitelaar, P., Sacaleanu, B. Ranking and selecting synsets by domain relevance. In: *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, NAACL 2001 Workshop, Carnegie Mellon University, Pittsburgh, 3-4 June 2001.

Buitelaar, P., Sacaleanu, B. Extending synsets with medical terms. In: *Proceedings of the First International WordNet Conference*, Mysore, India, January 21-25, 2002.

Burg, J. F. M., van de Riet, R. P. COLOR-X: Using knowledge from WordNet for conceptual modeling. Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Maryland, May 1998.

Burgun, A., Bodenreider, O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, 2001; 77-82.

Collins, A. M., Quillian, M. R. Retrieval time from semantic memory, *Journal of Verbal Learning and Verbal Behavior* 1969; 8: 240-248.

Cruse, D. A. *Lexical semantics*. Cambridge University Press, Cambridge, UK, 1986.

Cucchiarelli, A., Velardi, P. Automatic selection of class labels from a thesaurus for an effective semantic tagging of corpora. *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, 1997; 380-387.

Ely, J. W., Osheroff, J. A., Gorman, PN., Ebell, M. H., Chambliss, M. L., Pifer, E. A., Stavri, P. Z. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 2000; 321: 429-432.

Fellbaum, C. (ed.). *WordNet: an electronic lexical database.* MIT Press, Cambridge, MA, 1998.

Fellbaum, C. Distinguishing verb types in a lexical ontology. *Proceedings of the Second International Workshop on Generative Approaches to the Lexicon*. ISSCO, Geneva, 2003.

Fellbaum, C. English verbs as a semantic net. *International Journal of Lexicography* 1990; 3 (4): 278-301.

Fellbaum, C. On the semantics of troponymy. In: *The Semantics of Relationships: An Interdisciplinary Perspective*. R. Green, C. A. Bean, S. H. Myaeng (eds.), Dordrecht, Kluwer, 2002; 23-34.

Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J. Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.

Guha, R., Lenat, D., Pittman, K., Pratt, D., Shepherd, M. Cyc: A midterm report. *Communications of the ACP* 1990; 33 (8).

Harabagiu, S. M., Moldovan, DI. A marker propagation text understanding and inference system. J. H. Stewman (ed.), *Proceedings of the 9th Florida Artificial Intelligence Research Symposium*, Key West, 1996; 55-59.

Jacquemart, P., Zweigenbaum, P. Towards a medical question-answering system: a feasibility study. In: P. Le Beux and R. Baud (eds.), *Proceedings of Medical Informatics Europe*, IOS Press, Amsterdam, 2003; 463-468.

Kingsbury, P., Palmer, M. From TreeBank to PropBank. *Proceedings of 3rd International Conference on Language Resources and Evaluation* (LREC-2002), Las Palmas, Spain, 2002.

Lenat, D. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM* 1995; 38 (11).

Magnini, B., Strapparava, C. Using WordNet to improve user modelling in a web document recommender system. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.

McCray, A. T., Tse, T. Understanding search failures in consumer health information systems. *Proceedings of the American Medical Informatics Symposium* 2003: 430-4.

Medin, D. L., Atran, S. (eds.) *Folkbiology*. Cambridge, MA: MIT Press, 1999.

Miller, G. A. WordNet: a lexical database for English. *Comm ACM* 38, 11, November 1995; 39-41.

Patel, V. L., Arocha, J. F., Kushniruk, A. Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. *Journal of Biomedical Informatics*, 2002; 35(1): 8-16.

Pustejovsky, J. *The generative lexicon*. MIT Press, Cambridge, 1995.

Rosch, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology*, General 1975; 104: 192-253.

Rosch, E. On the internal structure of perceptual and semantic categories. *Cognitive Development and the Acquisition of Language*, T. E. Moore (ed.), Academic Press, New York, 1973.

Rosch, E. Principles of categorization. In: *Cognition and Categorization.* E. Rosch and B. B. Lloyd (eds.), Erlbaum, Hillsdale, NJ, 1978.

Slaughter, L. *Semantic relationships in health consumer questions and physicians' answers: a basis for representing medical knowledge and for concept exploration interfaces.* Doctoral dissertation, University of Maryland at College Park, 2002.

Smith, B., Köhler, J., Kumar, A. On the application of formal principles to life science data: a case study in the Gene Ontology. *Proceedings of DILS 2004* (Data Integration in the Life Sciences), (Lecture Notes in Computer Science), Berlin, Springer, 2004, in press.

Smith B., Rosse C. The role of foundational relations in the alignment of biomedical ontologies. *Proceedings of Medinfo*, San Francisco, 7-11 September, 2004.

Smith, C. A., Stavri, P. Z., Chapman, W. W. In their own words? A terminological analysis of e-mail to a cancer information service. *Proceedings of AMIA Symp*. 2002;: 697-701.

Tse, A., Soergel, D. Procedures for mapping vocabularies from non-professional discourse: a case study: in 'consumer medical vocabulary'. *Proceedings of the Annual Meeting of the American Society for Information*, 2003.

Tse, A. Y. *Identifying and characterizing a consumer medical vocabulary*. Doctoral dissertation, College of Information Studies, University of Maryland, College Park, Maryland, March 2003.

Turcato, D., Fass, D., Tisher, G., Popowich, F. Fully automatic bilingual lexical acquisition from EuroWordNet. *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.

Xiao, C., Rösner, D. Finding high-frequent synonyms of domain-specific verbs in the English sub-language of MEDLINE abstracts using WordNet. *Proc 2nd Global WordNet Conf* (GWC 2004), Brno, Czech Republic, December 2003; 242-247.

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., Boxwala, A. A. Characteristics of consumer terminology for health information retrieval: A formal study of use of a health information service. *Methods of Information in Medicine*, in press.